Historisches Textkorpus

von Elke Donalies und Ulrike Haß-Zumkehr

Seit Mitte 1992 gehört die Erforschung der neueren Geschichte des Deutschen satzungsgemäß zu den Aufgaben des Instituts für deutsche Sprache. Einige der bisherigen Arbeitsprojekte, vor allem das Deutsche Fremdwörterbuch und das Projekt zur deutschen Lehnwortbildung, besaßen von Anfang an eine diachronische Komponente. Die Lexikographinnen und Lexikographen wollen auch die jeweilige Vorgeschichte des heutigen Wortgebrauchs darstellen. Die vorhandenen sprachwissenschaftlichen Untersuchungen und Wörterbücher reichen dazu nicht aus; es müssen zeitgenössische Quellen herangezogen werden, die es gestatten, sich ein Bild von früheren Zuständen des Sprachgebrauchs zu machen.

In der Lexikographie benutzt man dazu traditionell sogenannte Belegkorpora, die aus Tausenden von Zetteln mit einzelnen Stichwörtern (eingebettet in Satzglieder oder vollständige, aber isolierte Sätze) und Quellenangaben bestehen. Das heißt, durch die Verzettelung werden zusammenhängende Texte in lexematische Atome zersplittert, die kein realistisches Bild des Wortgebrauchs mehr vermitteln. Deshalb bemüht man sich in neuerer Zeit mit Hilfe des Computers, die Belegsammlungen durch Korpora vollständiger Texte zu ergänzen.

Aber auch bei maschinellen Textkorpora darf man nicht vergessen, daß es sich nur um Momentaufnahmen eines kleinen, nämlich des schriftlich festgehaltenen Ausschnitts der sprachlichen Realität unserer Vorfahren handelt. Um dieser Realität so nah wie möglich zu kommen, muß die Textauswahl sprach- und literatursoziologische sowie dialekt- und textsortengeschichtliche Prinzipien zugrundelegen. Folgende Textsortenbereiche sind für das in der Abteilung »Historische Lexikologie und Lexikographie« geplante, umfangreiche Textkorpus vorgesehen: Zeitungen/Zeitschriften (ca. 30 %), biographische Texte (ca. 15 %), Sachliteratur (ca. 15 %), Texte des Kommunikationsbereichs Politik/Recht (ca. 15 %), Schöne und kulturelle »Höhenkamm«-Literatur (ca. 10 %), wissenschaftliche Texte (ca. 10 %), Texte der Fach- und Berufskommunikation (ca. 5 %). Die Bereiche sind in weitere Textsorten- und thematische Gruppen unterteilt.

Der zeitliche Rahmen umfaßt das 18. und 19. Jahrhundert, wobei die sprachgeschichtlich relevanten Zäsuren einerseits eher bei 1650 als bei 1700 und andererseits bei 1918 statt 1900 liegen. Es ist sowohl unter sprachkommunikativem Gesichtspunkt als auch aus Gründen der Quellenverfügbarkeit durchaus legitim, gewisse Zeitspannen, z.B. die Jahre um 1848, etwas stärker zu berücksichtigen als andere. Die Organisation des Projekts erlaubt ständige Kontrolle und spätere Korrekturen sowohl der thematisch-kommunikativen als auch der zeitlichen Gewichtung.

Vielfältige Nutzungsmöglichkeiten

Das Historische Textkorpus wird von unterschiedlichen wissenschaftlichen Fragestellungen aus genutzt werden können. Was den Kollegen aus den gegenwartssprachbezogenen Abteilungen des IDS im Hinblick auf die in Mannheim bereits seit langem installierten Computerkorpora zum Sprachgebrauch nach 1945 recht ist, ist den Sprachhistorikern billig. Auch sie wollen Textquellen computergestützt analysieren können, also z.B. wissen, wie oft ein Wort in welchen Texten oder Gattungen vorkam und mit welchen anderen Wörtern es vorzugsweise kombiniert wurde – hier können auch Beziehungen zwischen Wörtern unabhängig davon untersucht werden, wie weit sie in einem Text auseinanderliegen.

Abgefragt werden kann auch, um weitere Nutzungsmöglichkeiten aufzuzeigen:

- die Verwendung von Namen für Personen und geographische Einheiten, für nationale oder ethnische Gruppen (z.B. *Indianer, Deutsche*), für Bücher, Musikwerke, Bühnenstücke usw.,
- standardisierte Formulierungen, die den semantischen Wandel von Leitbegriffen wie ›Freiheit‹ oder ›Arbeit‹ mitbedingt haben, in denen sich die Etablierung einer öffentlichen Meinung, die Einführung telegraphischer Nachrichtendienste, die Einflüsse bestimmter Literaturgattungen oder die Zensurbestimmungen widerspiegeln,
- Personengruppen, die in Preußen oder in St. Petersburg unter *Ausländer* subsumiert wurden,
- Handlungsverben, die mit öffentlicher Meinung verbunden wurden.

Selbstverständlich kann das Korpus auch viele orthographie- und syntaxgeschichtliche Forschungen unterstützen, etwa zur früheren Funktion des Konjunktivs, zu Flexion und Wortstellung.

Neue Anforderungen

Drei entscheidende Neuerungen soll das Historische Korpus des IDS also bieten: Erstens soll es ganze Texte und nicht nur isolierte Sätze oder Satzteile enthalten. Zweitens soll es dem Stand der sprachgeschichtlichen Forschung entsprechen, die sich in den letzten Jahrzehnten zunehmend sozialgeschichtlichen Aspekten des Sprachgebrauchs geöffnet hat. Und drittens wird mit dem Einsatz der Computertechnik ein Sprung aus den staubigen Bücherregalen und unübersichtlichen Zettelkästen ins moderne Informationszeitalter getan.

Um Antworten auf die oben skizzierten und viele andere Fragen zu bekommen, müssen die Texte in bestimmter Weise aufbereitet, mit Markierungen versehen und an ein Abfrageprogramm angeschlossen werden. Bis Texte wie die von Friedrich dem Großen, Joh. Friedrich Blumenbach, Therese Huber und ihrem Mann Georg Forster, Alexander von Humboldt, Otto von Bismarck, Rudolf Virchow, Heinrich Schliemann, Fanny Lewald oder Henriette Davidis in der genannten Weise zur Verfügung stehen, sind noch viele Probleme gemeinsam von Linguisten und EDV-Fachleuten im IDS zu lösen:

Schon aus der aus heutiger Sicht abweichenden Flexion (z.B. *er frug dem Publico*) ergeben sich neue Anforderungen an das Lemmatisierungsprogramm, eines der unsichtbaren Herzstücke des Textaufbereitungs- und Abfrageprogramms. Wieviel leichter wäre es etwa, wenn die heute heiß umkämpfte Orthographieregelung schon 300 Jahre früher durchgeführt worden wäre. Wie soll der Computer begreifen, daß *Freyheit, Vraiheyt, freihayt* und *vriiheit, wohl* und *wol, Kultur* und *Cultur* nur Schreibvarianten einunddesselben Worts sind, die er alphabetisch zusammenzuordnen hat? Die Groß- und Kleinschreibung gilt heute als unantastbar, aber welche Freiheiten besaßen unsere Vorfahren, je nach Bedeutungsunterschied *Herr, HErr* oder *HERR, ER* oder *Er* zu schreiben! Solche und andere Besonderheiten in der Schreibweise dürfen bei der Texterfassung keinesfalls eingeebnet, sondern müssen der Authentizität wegen beibehalten werden oder doch rekonstruierbar sein.

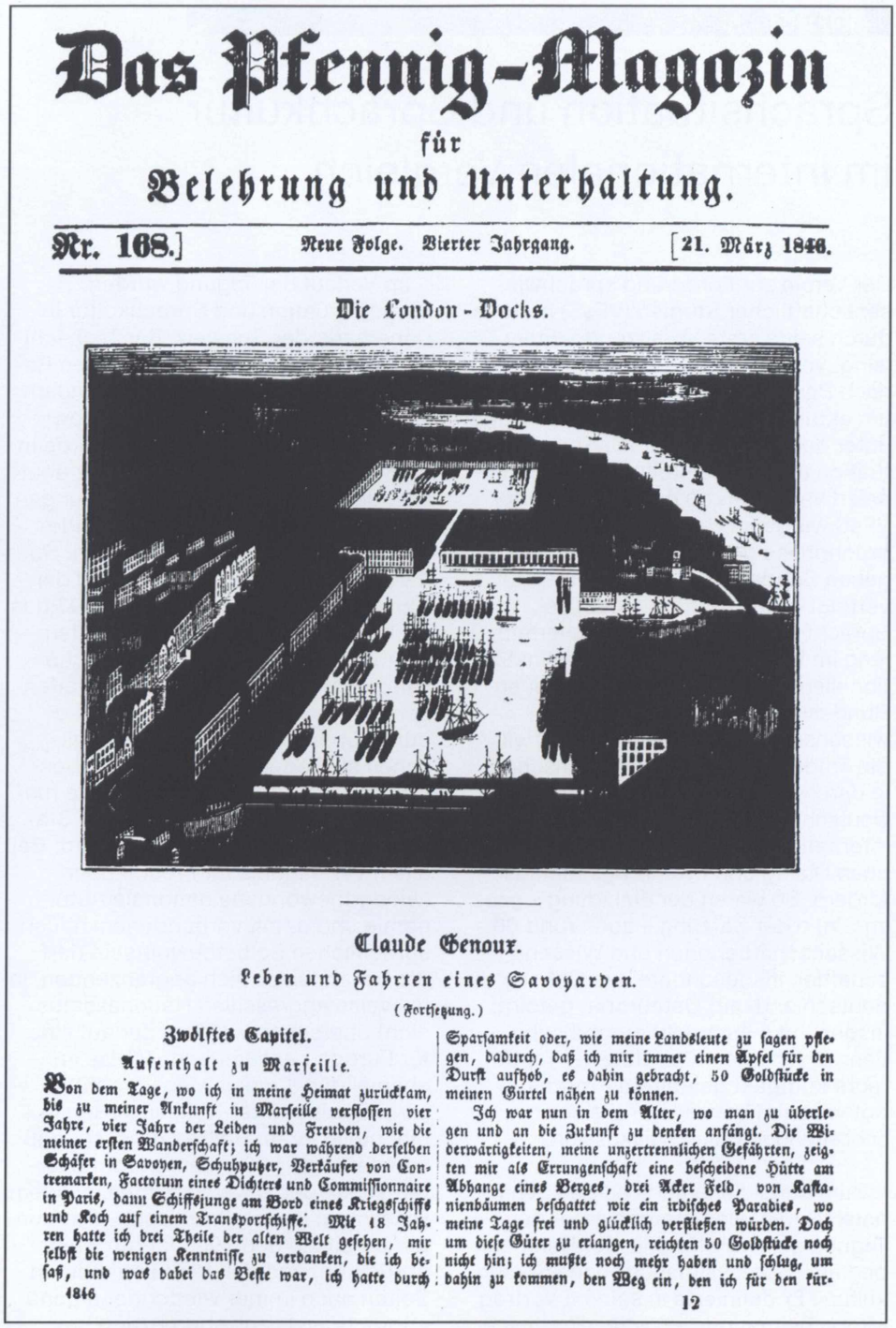
Die Liste dieser und weiterer in die originalen Texte (s. Abb. 1) einzufügenden Markierungen, die übrigens oft am Anfang und am Ende des zu Markierenden stehen müssen (s. Abb. 2), ist weitaus länger als bei gegenwartssprachigen Computerkorpora. Obligatorisch für alle Korpora (unabhängig von der Zeitstufe) sind die bibliographisch bzw. philologisch wichtigen Phänomene wie Seitenzahl, Überschrift, Absatz, Satz, typographischer Wechsel. Schwierigkeiten machen des weiteren nicht nur Randglossen, Fußnoten, tabellenartige Listen oder Bildunterschriften; diskutiert werden muß bereits darüber, wo ein Satz oder Absatz anfängt und aufhört. Eher fakultativ, aber dafür besonders vielversprechend bei späteren Abfragen sind die Markierungen für Namen, Abkürzungen, Zitate oder fremdsprachliche Einschübe (z.B. *»Lernst auch brav? Wieviel verba anomala sind?« »Ich weiß es nicht.«* Jung-Stilling 1777).

»Linguistische Bauchschmerzen«

Probleme ergeben sich vor allem bei der Bestimmung von Namen, etwa in Zusammensetzungen (wie *Mozartabend, Faust-Darsteller*), in Ableitungen (wie *britische Regierung, Münchener Bürger, russifizieren, göthisch*), im metaphorischen Gebrauch von Namen (z.B. bei *Westentaschencasanova*) und bei der Verwendung von Appellativa mit namenartiger Funktion (wie *die Republik* für Frankreich). Ganz zu schweigen von Wörtern in griechischer, kyrillischer und hebräischer Schrift oder von den vielen heute kaum reproduzierbaren Sonderzeichen mit zweifellos sprachlicher Funktion (z.B. das Pfund-Zeichen oder das beliebte Zeigehändchen).

Beim Aufbau eines historischen Textkorpus muß man zu fast allen linguistischen Grundfragen Stellung beziehen. Hier kann man sich nicht mit einer problematisierenden Darstellung zufriedengeben, sondern muß pragmatische Lösungen finden – selbst wenn manchmal linguistische Bauchschmerzen bleiben.

Die Autorinnen sind wissenschaftliche Mitarbeiterinnen am Institut für deutsche Sprache.



Das Pfennig-Magazin
für
Belehrung und Unterhaltung.

Nr. 168.] Neue Folge. Vierter Jahrgang. [21. März 1848.

Die London-Docks.

Claude Genoux.

Leben und Fahrten eines Savoyarden.

(Fortsetzung.)

Zwölftes Capitel.

Aufenthalt zu Marseille.

Von dem Tage, wo ich in meine Heimat zurückkam, bis zu meiner Ankunft in Marseille verflossen vier Jahre, vier Jahre der Leiden und Freuden, wie die meiner ersten Wanderschaft; ich war während derselben Schäfer in Savoyen, Schuhputzer, Verkäufer von Contremarken, Factotum eines Dichters und Commissionnaire in Paris, dann Schiffsjunge am Bord eines Kriegsschiffs und Koch auf einem Transportschiffe. Mit 18 Jahren hatte ich drei Theile der alten Welt gesehen, mir selbst die wenigen Kenntnisse zu verdanken, die ich besaß, und was dabei das Beste war, ich hatte durch Sparsamkeit oder, wie meine Landsleute zu sagen pflegen, dadurch, daß ich mir immer einen Apfel für den Durst aufhob, es dahin gebracht, 50 Goldstücke in meinen Gürtel nähen zu können.

Ich war nun in dem Alter, wo man zu überlegen und an die Zukunft zu denken anfängt. Die Widerwärtigkeiten, meine unzertrennlichen Gefährten, zeigten mir als Errungenschaft eine bescheidene Hütte am Abhange eines Berges, drei Acker Feld, von Kastanienbäumen beschattet wie ein irdisches Paradies, wo meine Tage frei und glücklich verfließen würden. Doch um diese Güter zu erlangen, reichten 50 Goldstücke noch nicht hin; ich mußte noch mehr haben und schlug, um dahin zu kommen, den Weg ein, den ich für den kür-

1846 12

Abb. 1: Textvorlage im Original

```
DEMO - Ausschnitt aus dem "Pfennig-Magazin"

<snr>1<\snr><ues>Das Pfennig-Magazin für Belehrung und
Unterhaltung.<sen><aen>
<abk>Nr.<\abk> 168.<sen> Neue Folge.<sen> Vierter
Jahrgang.<sen> 21. März 1848.<sen><aen><\ues>
<bil>Die <gna>London-Docks<\gna>.<sen><aen><\bil>
<ues><pna>Claude Genoux<\pna>.<sen><aen>
Leben und Fahrten eines <grp>Savoyarden<\grp>.<sen><aen>
(Fortsetzung.)<sen><aen><\ues>
<ues>Zwölftes Capitel.<sen><aen>
Aufenthalt zu <gna>Marseille<\gna>.<\ues><sen><aen>
Von dem Tage, wo ich in meine Heimat zurückkam, bis
zu meiner Ankunft in <gna>Marseille<\gna> verflossen
vier Jahre, vier Jahre der Leiden und Freuden, wie die
meiner ersten Wanderschaft; ich war während derselben
Schäfer in <gna>Savoyen<\gna>, Schuhputzer, Verkäufer von
Contremarken, Factotum eines Dichters und Commissionnaire
in <gna>Paris<\gna>, dann Schiffsjunge am Bord eines
Kriegsschiffs und Koch auf einem Transportschiffe.<sen>
```

Abb. 2: Markierte, computerlesbare Fassung